Code Summarization: Results Achieved with Zero-shot + Instructions

Table 1: Number (#) and percentage (%) of instances for which the LLMs did not manage to output a valid judgment.

	Java		**Python**	
	#	%	#	%
CodeLlama 7B	1	0.17%	0	0.00%
CodeLlama 13B	23	3.87%	2	0.35%
CodeLlama 34B	2	0.34%	13	2.28%
GPT-3.5-turbo	0	0.00%	0	0.00%
GPT-4-turbo	0	0.00%	0	0.00%

	CL 7B	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**	**Human Written**	**Own *vs* LLMs**	**Own *vs* LLMs \ F**	**Own *vs* Human**
Content Adequacy									
human	2.89	3.59	3.73	4.11	4.67	2.97	-	-	-
CL 7B	1.34	0.61	0.40	0.26	-0.42	1.24	*** (L)	*** (L)	(N)
CL 13B	1.04	0.36	0.23	-0.17	-0.67	0.91	* (S)	*** (M)	** (S)
CL 34B	1.62	1.11	0.80	0.77	0.30	1.70	(N)	(N)	*** (M)
GPT 3.5	1.13	0.55	0.43	-0.08	-0.55	1.02	*** (S)	*** (M)	*** (L)
GPT 4	0.35	-0.23	-0.16	-0.30	-0.56	0.15	*** (S)	*** (M)	*** (L)
Conciseness									
human	4.53	4.74	4.57	4.92	4.82	4.80	-	-	-
CL 7B	-0.38	-0.70	-0.55	-0.70	-0.74	-0.71	** (S)	** (S)	** (S)
CL 13B	-1.72	-1.86	-1.74	-2.10	-1.91	-2.04	(N)	(N)	(S)
CL 34B	-0.52	-0.42	-0.63	-0.35	-0.36	-0.88	(N)	* (S)	(N)
GPT 3.5	0.25	0.01	0.14	0.07	0.03	-0.17	(N)	(N)	** (S)
GPT 4	-0.25	-0.33	-0.20	-0.04	0.11	-0.24	*** (S)	*** (S)	*** (S)
Fluency&Understandability									
human	4.40	4.68	4.80	4.67	4.83	3.71	-	-	-
CL 7B	0.18	-0.12	-0.45	0.07	-0.18	1.06	** (S)	(N)	*** (M)
CL 13B	-0.52	-0.74	-0.88	-0.80	-0.88	0.10	(N)	(N)	*** (M)
CL 34B	0.36	0.13	-0.09	0.33	0.17	1.12	*** (S)	*** (S)	*** (L)
GPT 3.5	0.28	0.08	-0.08	0.08	0.01	0.84	(N)	(N)	*** (M)
GPT 4	0.28	0.08	-0.02	0.28	0.12	0.88	(N)	(N)	*** (M)

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 2: **(Java)** Code Summarization: Average of differences between the LLM judgments and the ground truth (*i.e.,* the median of the score given by the three human raters). Last three columns report adj. p-value and effect size when comparing the judgments each LLM gave to summaries it generated against those it gave when judging summaries (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

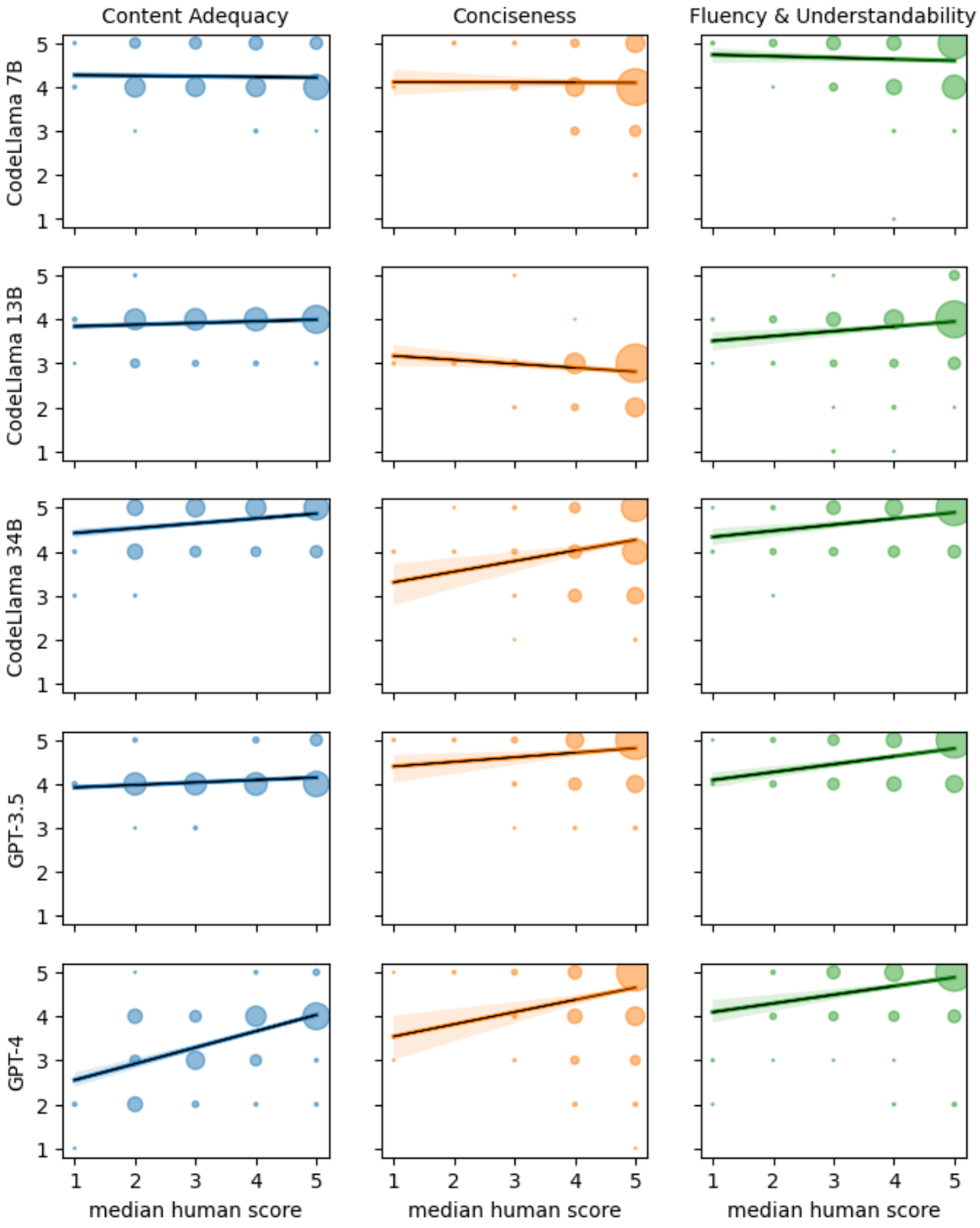

Figure 1: **(Java)** Code Summarization: Scatterplots relating human to LLM judgments.

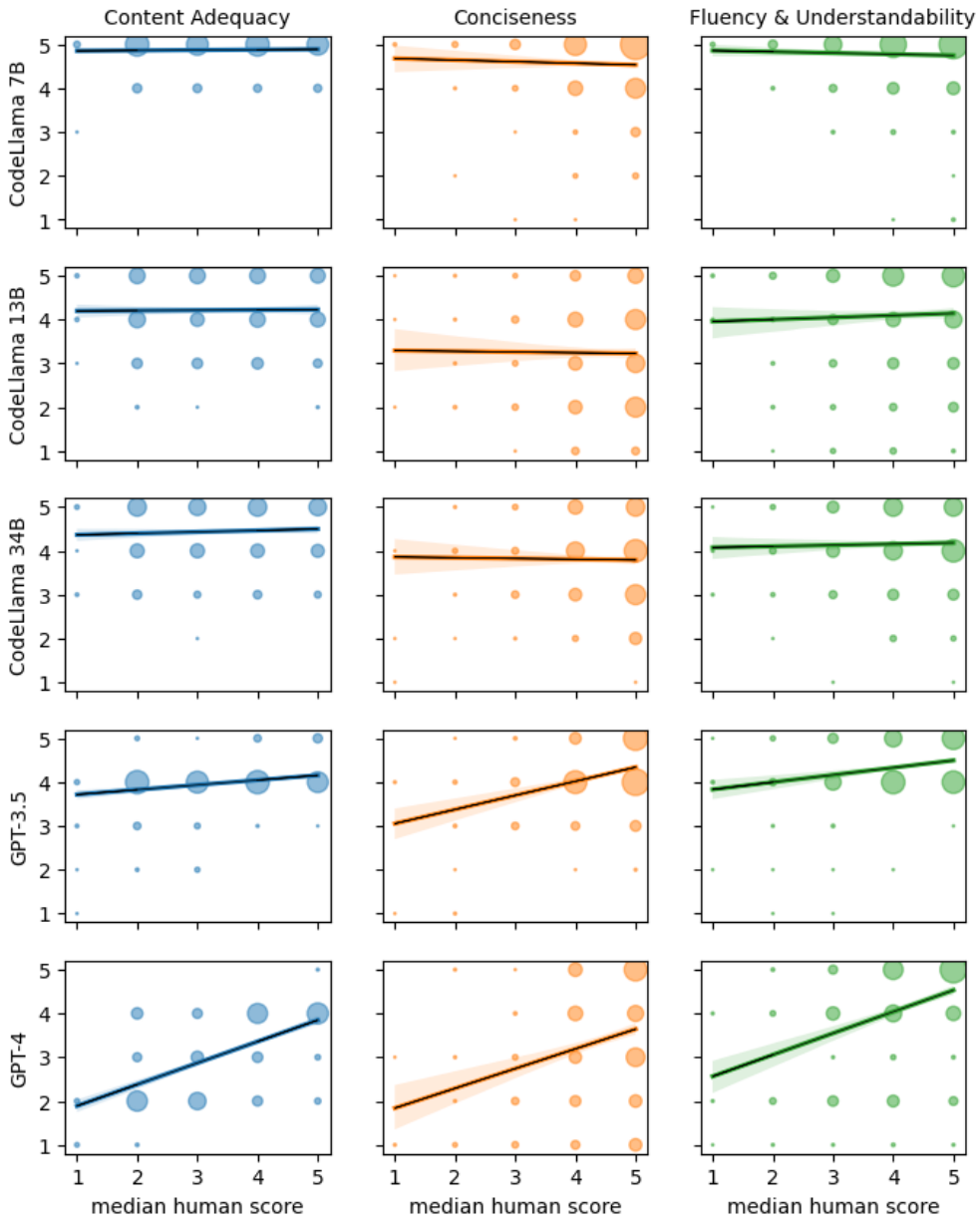

Figure 2: **(Python)** Code Summarization: Scatterplots relating human to LLM judgments.

	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
Content Adequacy									
human	2.95	3.05	2.78	3.99	4.54	2.93	-	-	-
CL 7B	1.95	1.83	2.10	0.87	0.33	1.95	*** (S)	*** (L)	(N)
CL 13B	1.26	1.28	1.44	0.21	-0.37	1.21	*** (S)	*** (L)	(N)
CL 34B	1.54	1.30	1.79	0.47	0.00	1.32	*** (M)	*** (L)	** (S)
GPT 3.5	1.02	0.85	1.08	0.07	-0.28	0.89	*** (S)	*** (L)	*** (L)
GPT 4	-0.07	-0.37	-0.27	-0.27	-0.55	-0.44	*** (S)	*** (S)	* (S)
Conciseness									
human	3.95	4.49	4.43	4.92	4.20	4.91	-	-	-
CL 7B	0.59	0.04	0.29	-0.46	0.36	-0.37	*** (S)	*** (S)	*** (M)
CL 13B	-0.65	-1.19	-1.28	-1.71	-1.05	-1.65	(N)	(N)	(S)
CL 34B	-0.08	-0.69	-0.55	-1.09	-0.29	-1.36	(N)	(N)	*** (M)
GPT 3.5	0.05	-0.40	-0.40	-0.35	0.00	-0.77	(N)	(N)	*** (S)
GPT 4	-1.17	-1.54	-1.83	-0.30	0.23	-1.98	*** (L)	*** (L)	*** (L)
Fluency&Understandability									
human	3.89	4.06	3.71	4.65	4.68	3.92	-	-	-
CL 7B	0.90	0.70	1.08	0.06	0.13	0.86	*** (S)	*** (L)	(N)
CL 13B	0.26	0.17	0.28	-0.73	-0.38	0.09	* (N)	*** (S)	(N)
CL 34B	0.31	0.08	0.45	-0.42	-0.49	0.15	*** (S)	*** (M)	(N)
GPT 3.5	0.27	0.26	0.53	-0.07	-0.17	0.39	* (S)	*** (S)	** (S)
GPT 4	-0.20	-0.33	-0.28	0.27	0.27	-0.06	** (S)	*** (S)	(S)

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 3: **(Python)** Code Summarization: Average of differences between the LLM judgments and the ground truth (*i.e.,* the median of the score given by the three human raters). Last three columns report adj. p-value and effect size when comparing the judgments each LLM gave to summaries it generated against those it gave when judging summaries (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

Code Summarization: Results Achieved with Automated Chain-of-Thought

Table 4: Number (#) and percentage (%) of instances for which the LLMs did not manage to output a valid judgment.

	Java		Python	
	#	%	#	%
CodeLlama 7B	47	7.91%	104	18.28%
CodeLlama 13B	28	4.71%	71	12.48%
CodeLlama 34B	9	1.52%	6	1.05%
GPT-3.5-turbo	0	0.00%	0	0.00%
GPT-4-turbo	0	0.00%	0	0.00%

	CL 7B	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**	**Human Written**	**Own *vs* LLMs**	**Own *vs* LLMs \ F**	**Own *vs* Human**
Content Adequacy									
human	2.89	3.59	3.73	4.11	4.67	2.97	-	-	-
CL 7B	0.43	-0.55	-0.46	-0.52	-1.09	0.20	*** (S)	*** (S)	(N)
CL 13B	0.49	-0.35	-0.33	-0.83	-1.05	0.05	(N)	(S)	(N)
CL 34B	1.00	0.22	0.23	-0.22	-0.80	0.88	(N)	** (S)	* (S)
GPT 3.5	1.84	1.08	1.05	0.74	0.29	1.47	(N)	*** (S)	*** (M)
GPT 4	0.22	-0.16	-0.02	-0.07	-0.05	0.22	(N)	(N)	* (S)
Conciseness									
human	4.53	4.74	4.57	4.92	4.82	4.80	-	-	-
CL 7B	0.25	-0.10	-0.03	-0.38	-0.14	-0.23	** (S)	*** (S)	** (S)
CL 13B	-1.50	-1.52	-1.37	-1.88	-1.67	-1.69	(N)	(N)	(N)
CL 34B	-0.22	-0.49	-0.31	-0.86	-0.75	-0.79	* (N)	** (S)	** (S)
GPT 3.5	0.37	0.07	0.06	-0.05	0.05	-0.04	** (S)	** (S)	(N)
GPT 4	-0.08	0.02	0.16	0.06	0.17	-0.07	(N)	(N)	** (S)
Fluency&Understandability									
human	4.40	4.68	4.80	4.67	4.83	3.71	-	-	-
CL 7B	0.35	-0.01	-0.21	0.10	-0.15	0.84	** (S)	(N)	*** (M)
CL 13B	-0.15	-0.61	-0.47	-0.54	-0.47	0.49	(N)	(N)	*** (M)
CL 34B	0.19	-0.04	-0.18	-0.11	-0.25	0.65	*** (S)	*** (S)	*** (L)
GPT 3.5	0.51	0.21	0.13	0.28	0.17	1.03	(N)	(N)	*** (M)
GPT 4	0.42	0.21	0.13	0.32	0.17	0.90	(N)	(N)	*** (M)

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 5: **(Java)** Code Summarization: Average of differences between the LLM judgments and the ground truth (*i.e.,* the median of the score given by the three human raters). Last three columns report adj. p-value and effect size when comparing the judgments each LLM gave to summaries it generated against those it gave when judging summaries (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

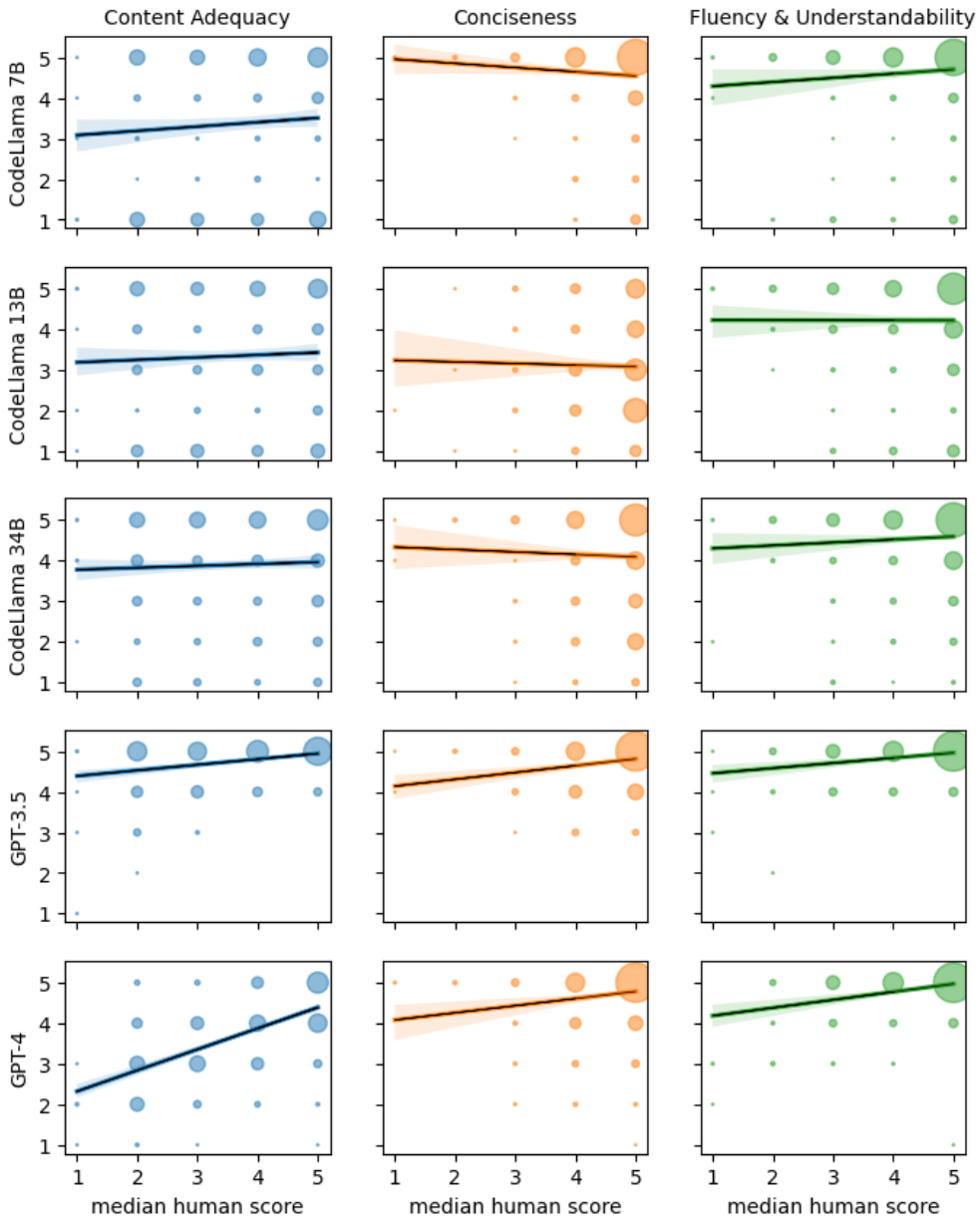

Figure 3: **(Java)** Code Summarization: Scatterplots relating human to LLM judgments.

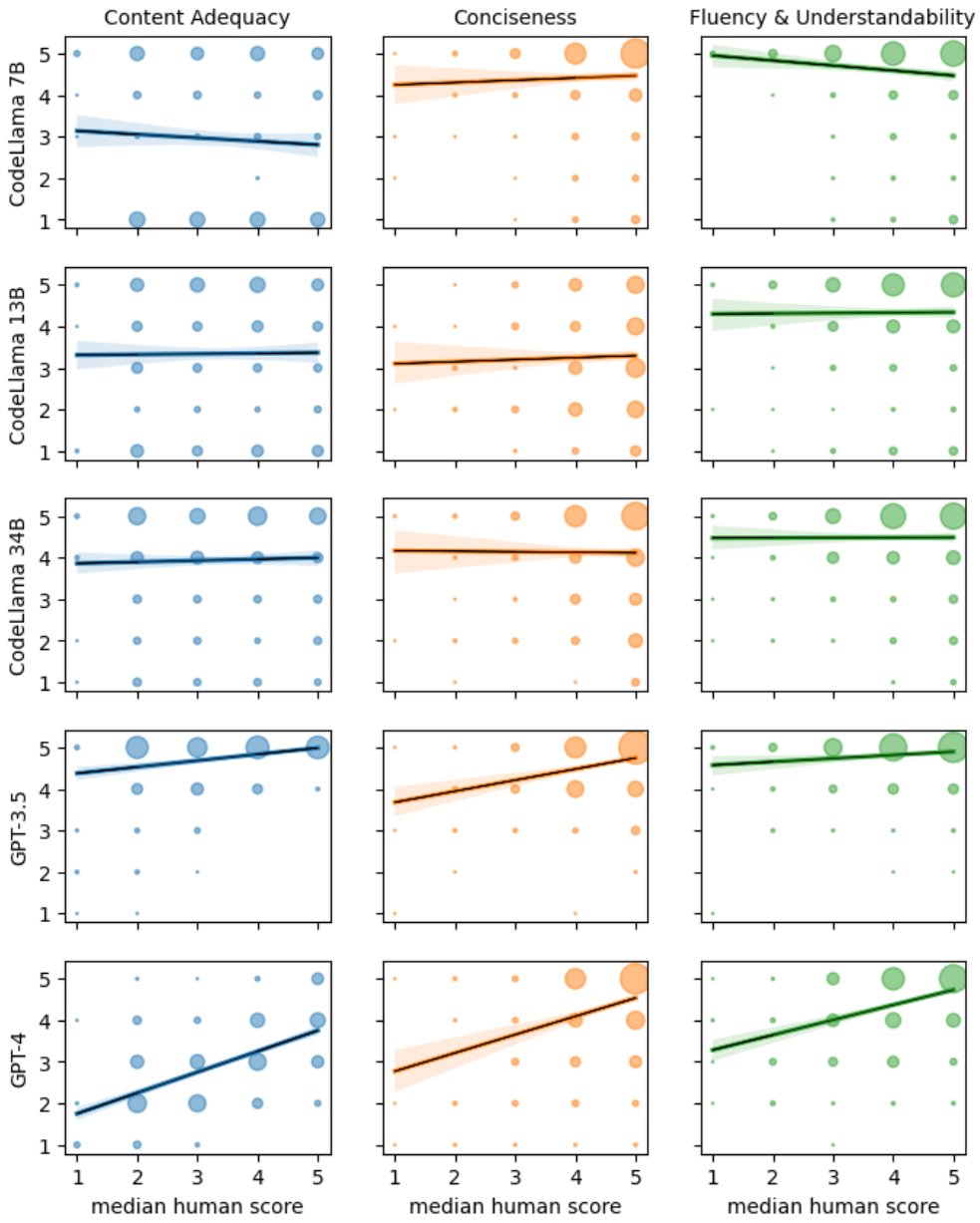

Figure 4: **(Python)** Code Summarization: Scatterplots relating human to LLM judgments.

	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
Content Adequacy									
human	2.95	3.05	2.78	3.99	4.54	2.93	-	-	-
CL 7B	-0.10	-0.01	-0.30	-1.04	-1.61	0.28	* (S)	*** (S)	(N)
CL 13B	0.41	0.14	0.41	-0.53	-1.23	0.43	(N)	** (S)	(N)
CL 34B	1.10	0.83	1.00	-0.12	-0.35	0.92	** (S)	*** (M)	(N)
GPT 3.5	1.87	1.67	1.86	0.87	0.43	1.52	*** (S)	*** (M)	*** (M)
GPT 4	-0.26	-0.49	-0.35	-0.72	-0.36	-0.46	(N)	(N)	(N)
Conciseness									
human	3.95	4.49	4.43	4.92	4.20	4.91	-	-	-
CL 7B	0.51	0.16	-0.03	-0.59	0.22	-0.42	*** (S)	*** (S)	*** (M)
CL 13B	-0.72	-1.33	-1.44	-1.60	-0.78	-1.46	(N)	(N)	(N)
CL 34B	0.15	-0.40	-0.26	-0.75	0.08	-0.99	(N)	(N)	** (S)
GPT 3.5	0.59	0.01	0.20	-0.08	0.35	-0.30	*** (S)	*** (S)	* (N)
GPT 4	-0.11	-0.53	-0.74	-0.12	0.77	-0.41	*** (L)	*** (L)	*** (L)
Fluency&Understandability									
human	3.89	4.06	3.71	4.65	4.68	3.92	-	-	-
CL 7B	0.79	0.57	0.96	-0.25	-0.20	0.66	*** (S)	*** (M)	(N)
CL 13B	0.33	0.07	0.70	-0.12	-0.25	0.38	(N)	* (S)	(N)
CL 34B	0.52	0.48	0.56	-0.02	-0.01	0.41	** (S)	*** (M)	(N)
GPT 3.5	0.99	0.72	1.05	0.25	0.26	0.80	*** (S)	*** (M)	*** (S)
GPT 4	0.21	0.14	0.31	0.24	0.28	0.40	(N)	(N)	(N)

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 6: **(Python)** Code Summarization: Average of differences between the LLM judgments and the ground truth (*i.e.,* the median of the score given by the three human raters). Last three columns report adj. p-value and effect size when comparing the judgments each LLM gave to summaries it generated against those it gave when judging summaries (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

Code Summarization: Results Achieved with Automated Chain-of-Thought + Instructions

Table 7: Number (#) and percentage (%) of instances for which the LLMs did not manage to output a valid judgment.

	Java		Python	
	#	%	#	%
CodeLlama 7B	35	5.89%	42	7.38%
CodeLlama 13B	12	2.02%	10	1.76%
CodeLlama 34B	9	1.52%	14	2.46%
GPT-3.5-turbo	0	0.00%	0	0.00%
GPT-4-turbo	0	0.00%	0	0.00%

	CL 7B	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**	**Human Written**	**Own *vs* LLMs**	**Own *vs* LLMs \ F**	**Own *vs* Human**
Content Adequacy									
human	2.89	3.59	3.73	4.11	4.67	2.97	-	-	-
CL 7B	1.59	0.91	0.91	0.38	0.02	1.44	*** (M)	*** (L)	(N)
CL 13B	1.18	0.11	-0.14	-0.56	-0.91	0.57	(N)	*** (S)	(N)
CL 34B	0.71	0.37	0.10	-0.25	-0.78	0.60	(N)	* (S)	* (S)
GPT 3.5	0.66	0.62	0.23	-0.09	1.13	** (S)	*** (S)	*** (L)	
GPT 4	-0.20	-0.37	-0.41	-0.09	-0.15	-0.08	(N)	(N)	(N)
Conciseness									
human	4.53	4.74	4.57	4.92	4.82	4.80	-	-	-
CL 7B	-0.38	-0.25	-0.05	-0.77	-0.37	-0.60	(N)	(N)	(N)
CL 13B	-1.11	-1.58	-1.53	-2.14	-1.52	-1.75	(N)	(N)	(N)
CL 34B	-0.69	-0.57	-0.65	-1.31	-0.79	-1.25	(N)	* (S)	* (S)
GPT 3.5	0.53	0.15	0.30	0.07	0.17	0.13	*** (S)	*** (S)	(N)
GPT 4	0.21	0.17	0.22	0.08	0.17	0.03	(N)	(N)	(N)
Fluency&Understandability									
human	4.40	4.68	4.80	4.67	4.83	3.71	-	-	-
CL 7B	-0.14	-0.19	-0.38	-0.42	-0.27	0.62	** (S)	** (S)	** (S)
CL 13B	-0.16	-0.67	-0.92	-0.85	-0.82	0.15	(N)	(N)	*** (S)
CL 34B	-0.21	-0.12	-0.47	-0.38	-0.41	0.35	(N)	(N)	*** (S)
GPT 3.5	0.57	0.29	0.15	0.32	0.18	1.11	(N)	(N)	*** (M)
GPT 4	0.24	-0.05	-0.11	0.12	0.02	0.59	(N)	(N)	*** (M)

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 8: **(Java)** Code Summarization: Average of differences between the LLM judgments and the ground truth (*i.e.,* the median of the score given by the three human raters). Last three columns report adj. p-value and effect size when comparing the judgments each LLM gave to summaries it generated against those it gave when judging summaries (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

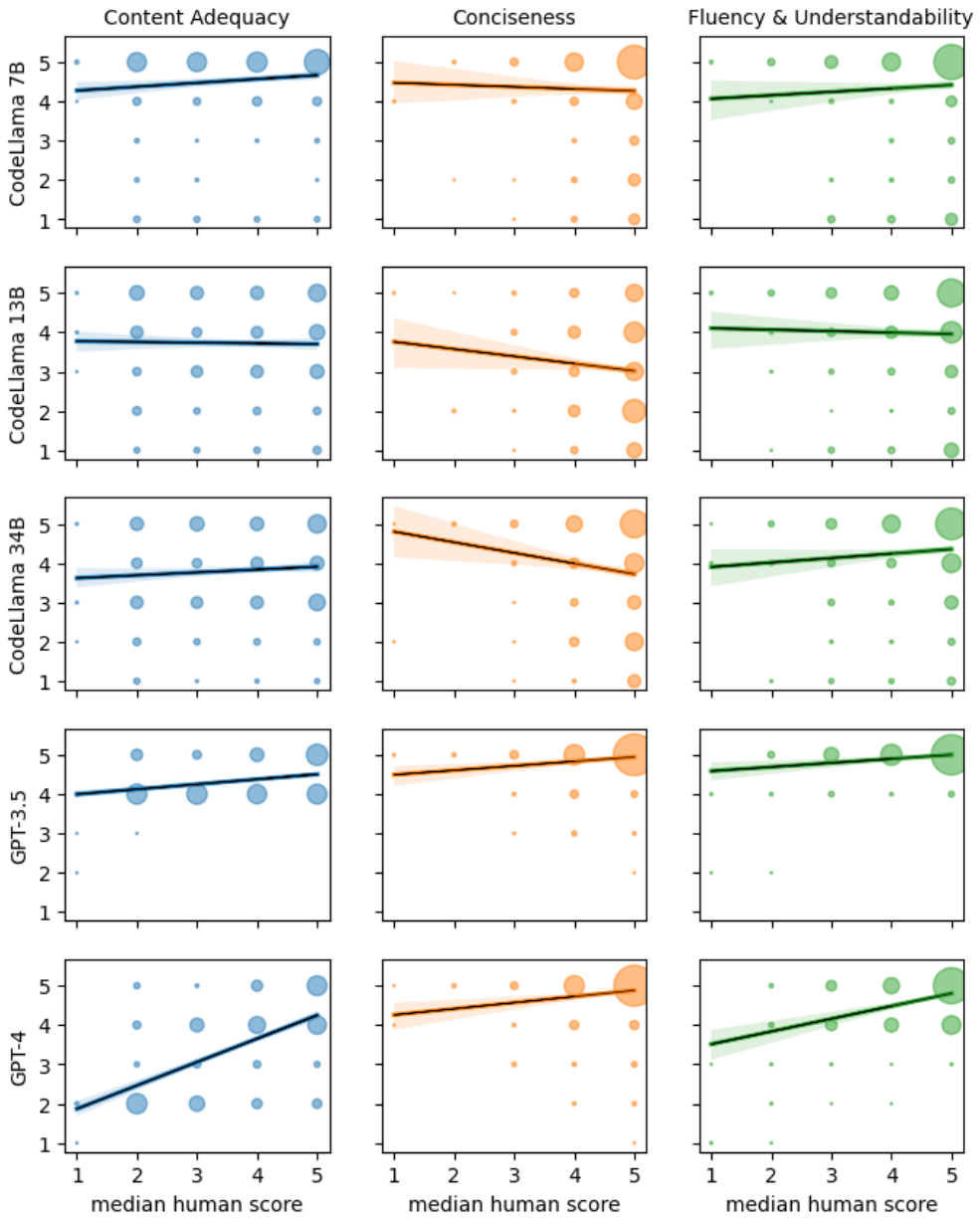

Figure 5: **(Java)** Code Summarization: Scatterplots relating human to LLM judgments.

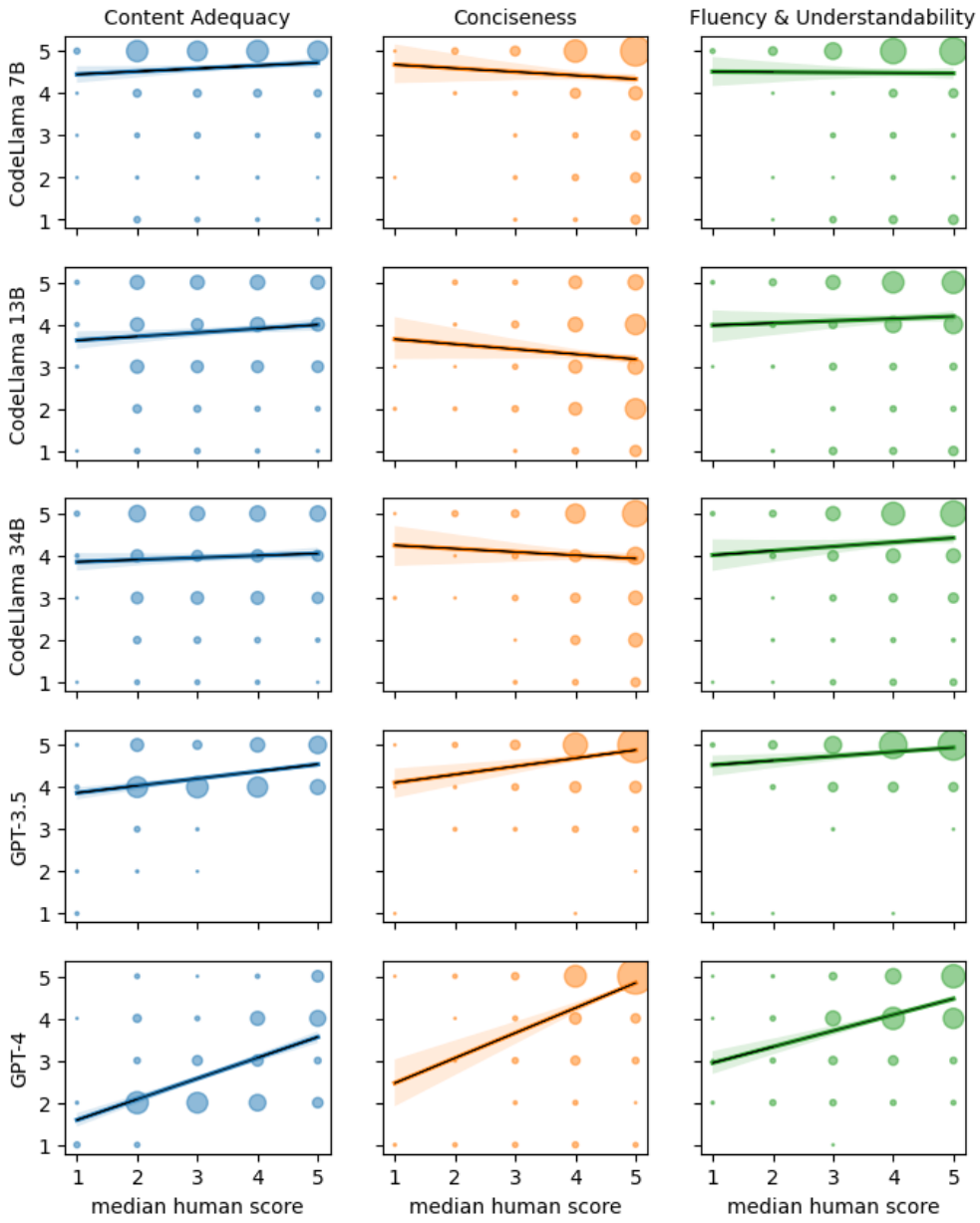

Figure 6: **(Python)** Code Summarization: Scatterplots relating human to LLM judgments.

	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
Content Adequacy									
human	2.95	3.05	2.78	3.99	4.54	2.93	-	-	-
CL 7B	1.61	1.68	1.88	0.62	0.16	1.49	** (S)	*** (L)	(N)
CL 13B	0.89	0.83	0.95	-0.12	-0.45	0.74	** (S)	*** (M)	(N)
CL 34B	1.14	0.78	1.02	0.12	-0.38	0.95	** (S)	*** (M)	(N)
GPT 3.5	1.33	1.09	1.33	0.33	0.16	1.12	*** (S)	*** (M)	*** (L)
GPT 4	-0.54	-0.76	-0.56	-0.90	-0.36	-0.63	(N)	(N)	(N)
Conciseness									
human	3.95	4.49	4.43	4.92	4.20	4.91	-	-	-
CL 7B	0.28	-0.11	0.00	-0.48	0.35	-0.62	** (S)	** (S)	*** (M)
CL 13B	-0.82	-1.16	-1.20	-1.69	-0.71	-1.88	(N)	(N)	** (S)
CL 34B	0.01	-0.65	-0.68	-0.82	-0.08	-0.93	(N)	(N)	(N)
GPT 3.5	0.80	0.14	0.35	0.02	0.59	-0.12	*** (M)	*** (S)	* (N)
GPT 4	0.01	-0.22	-0.31	0.06	0.76	-0.06	*** (L)	*** (L)	*** (L)
Fluency&Understandability									
human	3.89	4.06	3.71	4.65	4.68	3.92	-	-	-
CL 7B	0.48	0.59	0.74	-0.18	-0.05	0.45	* (S)	*** (M)	(N)
CL 13B	0.22	0.00	0.42	-0.36	-0.28	-0.02	(N)	(S)	(N)
CL 34B	0.42	0.20	0.43	-0.18	-0.11	0.35	* (S)	*** (S)	(N)
GPT 3.5	1.02	0.73	1.07	0.28	0.29	0.80	*** (S)	*** (M)	*** (S)
GPT 4	-0.05	-0.18	0.19	-0.11	-0.04	0.12	(N)	(N)	(N)

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 9: **(Python)** Code Summarization: Average of differences between the LLM judgments and the ground truth (*i.e.,* the median of the score given by the three human raters). Last three columns report adj. p-value and effect size when comparing the judgments each LLM gave to summaries it generated against those it gave when judging summaries (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.